CAL 282

Probability & Statistics

2013-2014 SPRING TERM

1st Week

Assist.Prof.Dr. Yılmaz BAYAR

Introduction

- Is everything on this planet determined by randomness? This question is open to philosophical debate. What is certain is that every day thousands and thousands of engineers, scientists, business persons, manufacturers, and others are using tools from probability and statistics.
- The theory and practice of probability and statistics were developed during the last century and are still actively being refined and extended.
- **Probability theory (olasılık teorisi)** is the study of the mathematical rules that govern **random events**.
- **A random event (rasgele olay)** is an event in which we do not know the outcome without observing it.
- Probability tells us what we can say about such events, given our assumptions about the possible outcomes.
- **Statistics (istatistik) is the application of probability to the collection, analysis, and description of random data. Statistics** can be described as the study of how to make inference and decisions in the face of uncertainty and variability.
- The discipline of **statistics** teaches us how to make intelligent judgments and informed decisions in the presence of uncertainty and variation.
- Without uncertainty or variation, there would be little need for statistical methods or statisticians.

Introduction

- You may be familiar with **probability** and **statistics** through radio, television, newspapers, and magazines.

Ex: Eating 10 grams of fiber a day reduces the risk of heart attack by 14%. (Source: Archives of Internal Medicine, Reader's Digest.)

- Statistics is used in almost all fields of human endeavor (Sports, public health, education etc.)
- Furthermore, statistics is used to analyze the results of surveys and as a tool in scientific research to make decisions based on controlled experiments. Other uses of statistics include operations research, quality control, estimation, and prediction.
- **Statistics is the science of conducting studies to collect, organize, summarize, analyze, and draw conclusions from data.**

Introduction

- **Applications of Probability and Statistics:**

- **Computer Science:**
  Machine Learning
  Data Mining
  Simulation
  Image Processing
  Computer Vision
  Computer Graphics
  Visualization
  Software Testing
  Algorithms

  Electrical Engineering:
  Signal Processing
  Telecommunications
  Information Theory
  Control Theory
  Instrumentation, Sensors
  Hardware/Electronics Testing

- General:
  Gambling (not recommended)
  Politics
  Demographics
  Economics

  Stock Market Analysis
  Sports
  Medicine
  All Sciences!!

Introduction

- **Students study statistics for several reasons:**
- You must be able to read and understand the various statistical studies performed in your fields.
- You may be called on to conduct research in your field, since statistical procedures are basic to research.
- You can also use the knowledge gained from studying statistics to become better consumers and citizens.

Statistical Terms

- **Variable (değişken):** A characteristic or attribute that can assume different values.
- **Data (veri):** Values (measurements or observations) that the variables can assume.
- **Random variables (rasgele değişken):** Variables whose values are determined by chance.
- **Deterministic variable (deterministik değişken):** A variable representing a deterministic relation

Ex: Suppose that an insurance company studies its records over the past several years and determines that, on average, 3 out of every 100 automobiles the company insured were involved in accidents during a 1-year period.

Although there is no way to predict the specific automobiles that will be involved in an accident (random occurrence), the company can adjust its rates accordingly, since the company knows the general pattern over the long run.

- **Data set (veri seti):** A collection of data values
- **Data value or datum (veriler):** Each value in the data set

Branches of Statistics

- **Descriptive statistics (betimleyici/tanımlayıcı istatistik)** consists of the collection, organization, summarization, and presentation of data.
- In descriptive statistics the statistician tries to describe a situation.
- The origin of descriptive statistics can be traced to data collection methods used in censuses taken by the Babylonians and Egyptians between 4500 and 3000 B.C.
- In addition, the Roman Emperor Augustus (27 B.C.—A.D. 17) conducted surveys on births and deaths of the citizens of the empire, as well as the number of livestock each owned and the crops each citizen harvested yearly.

Branches of Statistics

- **Inferential statistics (çıkarımsal istatistik)** consists of generalizing from samples to populations, performing estimations and hypothesis tests, determining relationships among variables, and making predictions.
- Inferential statistics originated in the 1600s, when John Graunt published his book on population growth, Natural and Political Observations Made upon the Bills of Mortality. About the same time, another mathematician/astronomer, Edmund Halley, published the first complete mortality tables. (Insurance companies use mortality tables to determine life insurance rates.)
- The statistician tries to make inferences from samples to populations.
- Inferential statistics uses **probability**,
- **Probability is the chance of an event occurring.**
- **A population (anakütle/evren)** consists of all subjects (human or otherwise) that are being studied.

Branches of Statistics

- Most of the time, due to the expense, time, size of population, medical concerns, etc., it is not possible to use the entire population for a statistical study; therefore, researchers use samples.
- **A sample (örnek/örneklem)** is a group of subjects selected from a population.
- **Hypothesis testing (hipotez testi):** An area of inferential statistics, a decision-making process for evaluating claims about a population, based on information obtained from samples.
- For example, a researcher may wish to know if a new drug will reduce the number of heart attacks in men over 70 years of age. For this study, two groups of men over 70 would be selected. One group would be given the drug, and the other would be given a placebo (a substance with no medical benefits or harm). Later, the number of heart attacks occurring in each group of men would be counted, a statistical test would be run, and a decision would be made about the effectiveness of the drug.

Branches of Statistics

- Statisticians also use statistics to determine relationships among variables.
- **Example:** 187,783 men were observed over a period of 45 months. The death rate from lung cancer in this group of volunteers was 10 times as great for smokers as for nonsmokers.
- By studying past and present data and conditions, statisticians try to make predictions based on this information.

Variables and Types of Data

- Statisticians statisticians gain information about a particular situation by collecting data for random variables.
- Variables can be classified as **qualitative** or **quantitative**.
- **Qualitative variables (nitel değişkenler)** are variables that can be placed into distinct categories, according to some characteristic or attribute.
- For example, if subjects are classified according to gender (male or female), then the variable gender is qualitative. Other examples of qualitative variables are religious preference and geographic locations.
- **Quantitative variables (nicel değişkenler)** are numerical and can be ordered or ranked.
- For example, the variable age is numerical, and people can be ranked in order according to the value of their ages. Other examples of quantitative variables are heights, weights, and body temperatures.
- Quantitative variables can be further classified into two groups: **discrete** and **continuous**.
- **Discrete variables (ayrık/kesintili değişken)** assume values that can be counted.
- **Continuous variables (sürekli/kesintisiz değişken)** can assume an infinite number of values between any two specific values. They are obtained by measuring. They often include fractions and decimals.

Variables and Types of Data

- Variables can be classified by how they are categorized, counted, or measured.
- This type of uses **measurement scales**,
- Four common types of scales are used: **nominal**, **ordinal**, **interval**, and **ratio**.
- **Nominal level of measurement (nominal ölçüm düzeyi)** classifies data into mutually exclusive (nonoverlapping) categories in which no order or ranking can be imposed on the data.
- A sample of college instructors classified according to subject taught (e.g., English, history, psychology, or mathematics) is an example of nominal-level measurement.
- Classifying survey subjects as male or female is another example of nominal-level measurement.
- Classifying residents according to zip codes is also an example of the nominal level of measurement. Even though numbers are assigned as zip codes, there is no meaningful order or ranking.
- Political party (Democratic, Republican, Independent, etc.), religion (Christianity, Judaism, Islam, etc.), and marital status (single, married, divorced, widowed, separated).

Variables and Types of Data

- **Ordinal level of measurement (ordinal ölçüm düzeyi)** classifies data into categories that can be ranked; however, precise differences between the ranks do not exist.
- Data measured at this level can be placed into categories, and these categories can be ordered, or ranked.
- For example, from student evaluations, guest speakers might be ranked as superior, average, or poor. Floats in a homecoming parade might be ranked as first place, second place, etc.
- **Interval level of measurement (aralık ölçüm düzeyi)** ranks data, and precise differences between units of measure do exist; however, there is no meaningful zero.
- This level differs from the ordinal level in that precise differences do exist between units.
- IQ is an example of such a variable. There is a meaningful difference of 1 point between an IQ of 109 and an IQ of 110. Temperature is another example of interval measurement, since there is a meaningful difference of 1F between each unit, such as 72 and 73F.
- **There is no true zero**.
- For example, IQ tests do not measure people who have no intelligence. For temperature, 0F does not mean no heat at all.

Variables and Types of Data

- **Ratio level of measurement (oran ölçüm düzeyi)** possesses all the characteristics of interval measurement, and there exists a true zero.
- In addition, true ratios exist when the same variable is measured on two different members of the population.
- Examples of ratio scales are those used to measure height, weight, area, and number of phone calls received.
- **Ratio scales have differences between units** (1 inch, 1 pound, etc.) and a **true zero**.
- **Ratio scale contains a true ratio between values**.
- For example, if one person can lift 200 pounds and another can lift 100 pounds, then the ratio between them is 2 to 1. Put another way, the first person can lift twice as much as the second person.

Variables and Types of Data

- **There is not complete agreement among statisticians about the classification of data into one of the four categories.**
- For example, some researchers classify IQ data as ratio data rather than interval. Also, data can be altered so that they fit into a different category.
- For instance, if the incomes of all professors of a college are classified into the three categories of low, average, and high, then a ratio variable becomes an ordinal variable.

Data Collection and Sampling Techniques

- Data can be used to describe situations or events
- Data can be collected in a variety of ways.
- One of the most common methods is through the use of **surveys**.
- Surveys can be done by using a variety of methods. Three of the most common methods are the **telephone survey**, the **mailed questionnaire**, and the **personal interview**.
- Data can also be collected in other ways, such as **surveying records** or **direct observation of situations**.
- Researchers use **sample**s to collect data and information about a particular variable from a large population.
- Using **samples** saves time and money and in some cases enables the researcher to get more detailed information about a particular subject.
- To obtain samples that are unbiased—i.e., that give each subject in the population an equally likely chance of being selected—statisticians use four basic methods of sampling: **random**, **systematic**, **stratified**, and **cluster sampling**.

Data Collection and Sampling Techniques

- **Random sampling (rasgele/tesadüfi örnekleme):** Random samples are selected by using chance methods or random numbers.
- One such method is to number each subject in the population. Then place numbered cards in a bowl, mix them thoroughly, and select as many cards as needed. The subjects whose numbers are selected constitute the sample. Since it is difficult to mix the cards thoroughly, there is a chance of obtaining a biased sample.
- Statisticians use another method of obtaining numbers. They generate random numbers with a computer or calculator.
- **Systematic Sampling (sistematik örnekleme):** Systematic samples are obtained by numbering each subject of the population and then selecting every kth subject.
- For example, suppose there were 2000 subjects in the population and a sample of 50 subjects were needed. Since 2000/50=40, then k=40, and every 40th subject would be selected; however, the first subject (numbered between 1 and 40) would be selected at random.
- Suppose subject 12 were the first subject selected; then the sample would consist of the subjects whose numbers were 12, 52, 92, etc., until 50 subjects were obtained.

Data Collection and Sampling Techniques

- **Stratified Sampling (katmanlı örnekleme):** Stratified samples are obtained by dividing the population into groups according to some characteristic that is important to the study, then sampling from each group. **Samples within the strata should be randomly selected**.
- For example, suppose the president of a two-year college wants to learn how students feel about a certain issue. Furthermore, the president wishes to see if the opinions of the first-year students differ from those of the second-year students. The president will randomly select students from each group to use in the sample.
- **Cluster Sampling (küme örneklemesi):** Population is divided into groups called clusters by some means such as geographic area or schools in a large school district, etc. Then the researcher randomly selects some of these clusters and uses all members of the selected clusters as the subjects of the samples.

Data Collection and Sampling Techniques

- Suppose a researcher wishes to survey apartment dwellers in a large city. If there are 10 apartment buildings in the city, the researcher can select at random 2 buildings from the 10 and interview all the residents of these buildings.
- **Cluster sampling is used when the population is large or when it involves subjects residing in a large geographic area**.
- For example, if one wanted to do a study involving the patients in the hospitals in New York City, it would be very costly and time-consuming to try to obtain a random sample of patients since they would be spread over a large area. Instead, a few hospitals could be selected at random, and the patients in these hospitals would be interviewed in a cluster.

Random	Subjects are selected by random numbers.
Systematic	Subjects are selected by using every *k*th number after the first subject is randomly selected from 1 through *k*.
Stratified	Subjects are selected by dividing up the population into groups (strata), and subjects are randomly selected within groups.
Cluster	Subjects are selected by using an intact group that is representative of the population.

Observational and Experimental Studies

- In an **observational study (gözlemsel çalışma)**, the researcher merely observes what is happening or what has happened in the past and tries to draw conclusions based on these observations.
- For example, data from the Motorcycle Industry Council (USA TODAY) stated that "Motorcycle owners are getting older and richer." Data were collected on the ages and incomes of motorcycle owners for the years 1980 and 1998 and then compared. The findings showed considerable differences in the ages and incomes of motorcycle owners for the two years.
- In this study, the researcher merely observed what had happened to the motorcycle owners over a period of time. There was no type of research intervention.
- In an **experimental study (deneysel çalışma)**, the researcher manipulates one of the variables and tries to determine how the manipulation influences other variables.

Observational and Experimental Studies

- For example, a study conducted at Virginia Polytechnic Institute and presented in Psychology Today divided female undergraduate students into two groups and had the students perform as many sit-ups as possible in 90 sec.
- The first group was told only to "Do your best," while the second group was told to try to increase the actual number of sit-ups done each day by 10%.
- After four days, the subjects in the group who were given the vague instructions to "Do your best" averaged 43 sit-ups, while the group that was given the more specific instructions to increase the number of sit-ups by 10% averaged 56 sit-ups by the last day's session.
- The conclusion then was that athletes who were given specific goals performed better than those who were not given specific goals.
- This study is an example of a statistical experiment since the researchers intervened in the study by manipulating one of the variables, namely, the type of instructions given to each group.

Observational and Experimental Studies

- Statistical studies usually include one or more **independent variables** and one **dependent variable**.
- **Independent/ explanatory variable (bağımsız/açıklayıcı değişken)** in an experimental study is the one that is being manipulated by the researcher.
- The resultant variable is called the **dependent/outcome variable (bağımlı değişken)**
- The outcome variable is the variable that is studied to see if it has changed significantly due to the manipulation of the independent variable.
- **A confounding variable (karıştırıcı/kirletici değişken)** is one that influences the dependent or outcome variable but was not separated from the independent variable.
- **Suspect samples** (small sample size, sampling technique)
- **Ambiguous averages** (mean, median, mode, and midrange)
- **Changing the subject**
- **Detached statistics**
- **Implied connections**
- **Misleading graphs**
- **Faulty survey questions**

Uses and Misuses of Statistics

- Statistical techniques can be used to describe data, compare two or more data sets, determine if a relationship exists between variables, test hypotheses, and make estimates about population characteristics.
- Misuse of statistical techniques are to sell products that don't work properly, to attempt to prove something true that is really not true, or to get our attention by using statistics to evoke fear, shock, and outrage.
- **"There are three types of lies—lies, damn lies, and statistics."**
- **"Figures don't lie, but liars figure."**